

Analytics in a day

Cloud analytics in the age of
self-service and data science

Steen Kjøng Kristensen

Senior Consultant, Business Intelligence
Accobat A/S

skp@accobat.com

<https://www.linkedin.com/in/supersteen/>



Agenda

Times are approximate and will be fluid with the class

Workshop Instruction & Discussion

- 9:00 – 9:30 Harness the power of analytics
- 9:30 – 10:30 Experience a new class of analytics with Azure Synapse
- 10:30 – 10:45 Get started with your first analytics project

Break for 15 min

Hands-on lab

- 11:00 – 13:00 Hands-on lab using Azure Synapse and Power BI

Harness the power of analytics

Section 1

All businesses are data businesses

Section 2

The cloud for modern analytics

Section 3

The paradox of analytics

Section 4

The analytics continuum



Section 1

All businesses are data businesses

Today, all businesses are data businesses

Data is the lifeblood of modern work



Data makes business possible

Data is crucial to businesses of all sizes

Securing data is mandatory

All data businesses need to be analytic businesses

Without analytics data is a cost center,
not a resource



**Businesses need to
“know thyself”**

**Workers are empowered
with intelligent insights**

**Augmented intelligence
sees patterns we cannot**

Analytic businesses need to evolve data science

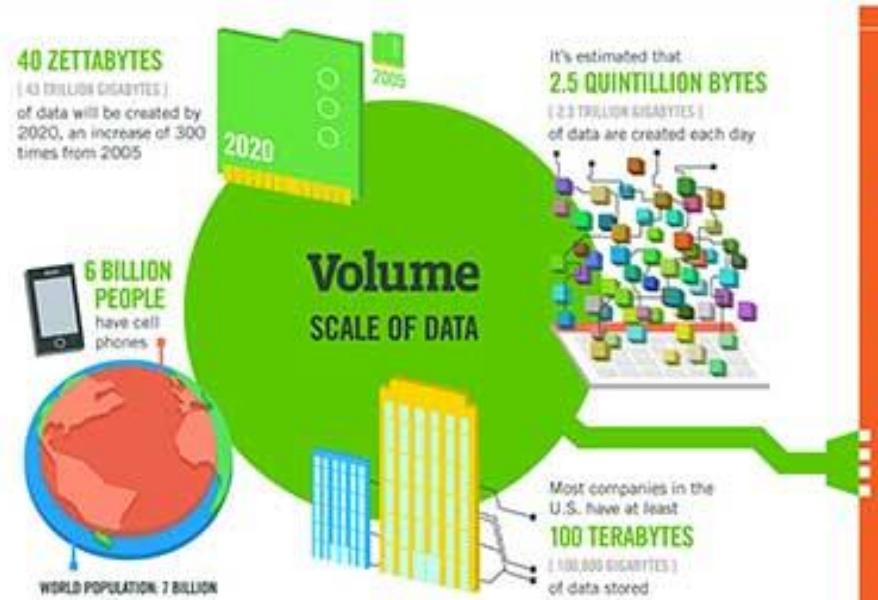
Every business has opportunities to make analytics faster, easier, and more insightful



Data science empowers us to go farther

AI/ML creates smarter organizations

Businesses need scalable analytics



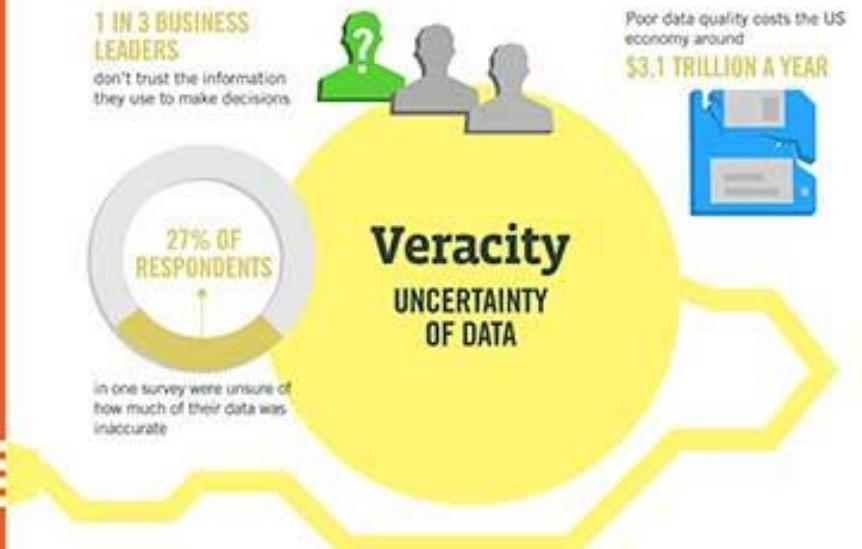
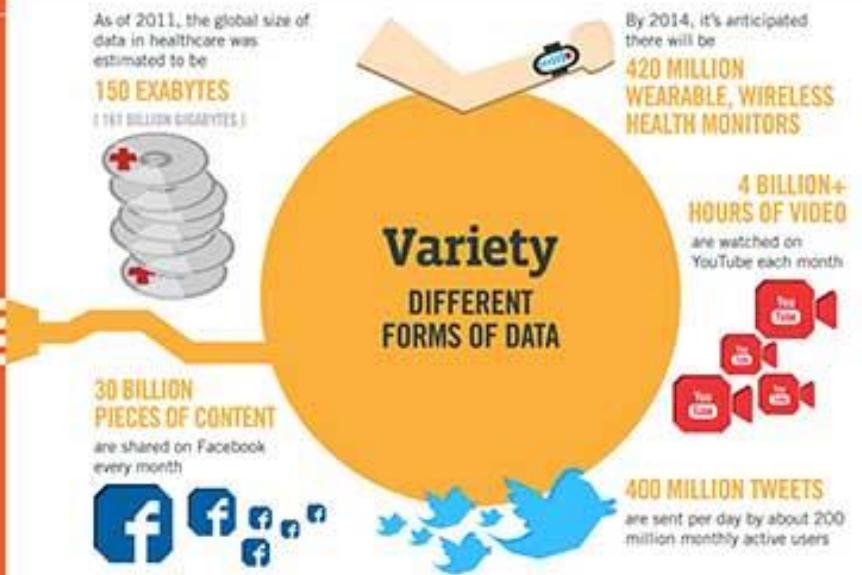
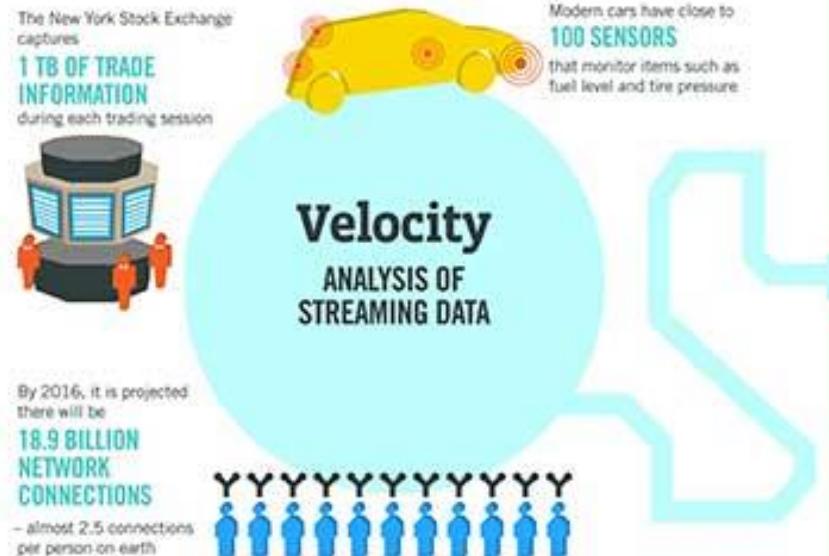
The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: Volume, Velocity, Variety and Veracity.

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS will be created globally to support big data, with 1.9 million in the United States.



Section 2

The cloud for modern analytics

Modern businesses succeed in the cloud

The cloud is the default environment for new technology initiatives



Applications are increasingly cloud-native

Collaboration platforms and file-sharing are transforming

Hybrid businesses are replacing on-premises data management

A cloud analytics platform is an economic breakthrough

Price, performance, and agility



High infrastructure costs are avoided with cloud-based analytics platforms

Iterating designs and handling diverse data becomes more efficient

Elastic scaling makes world-class performance available to all users

A cloud analytics platform is the hub for all data models

Structured, unstructured, and streaming data
integrated in a single, scalable, environment



**Workloads can be
stored, managed and
provisioned together**

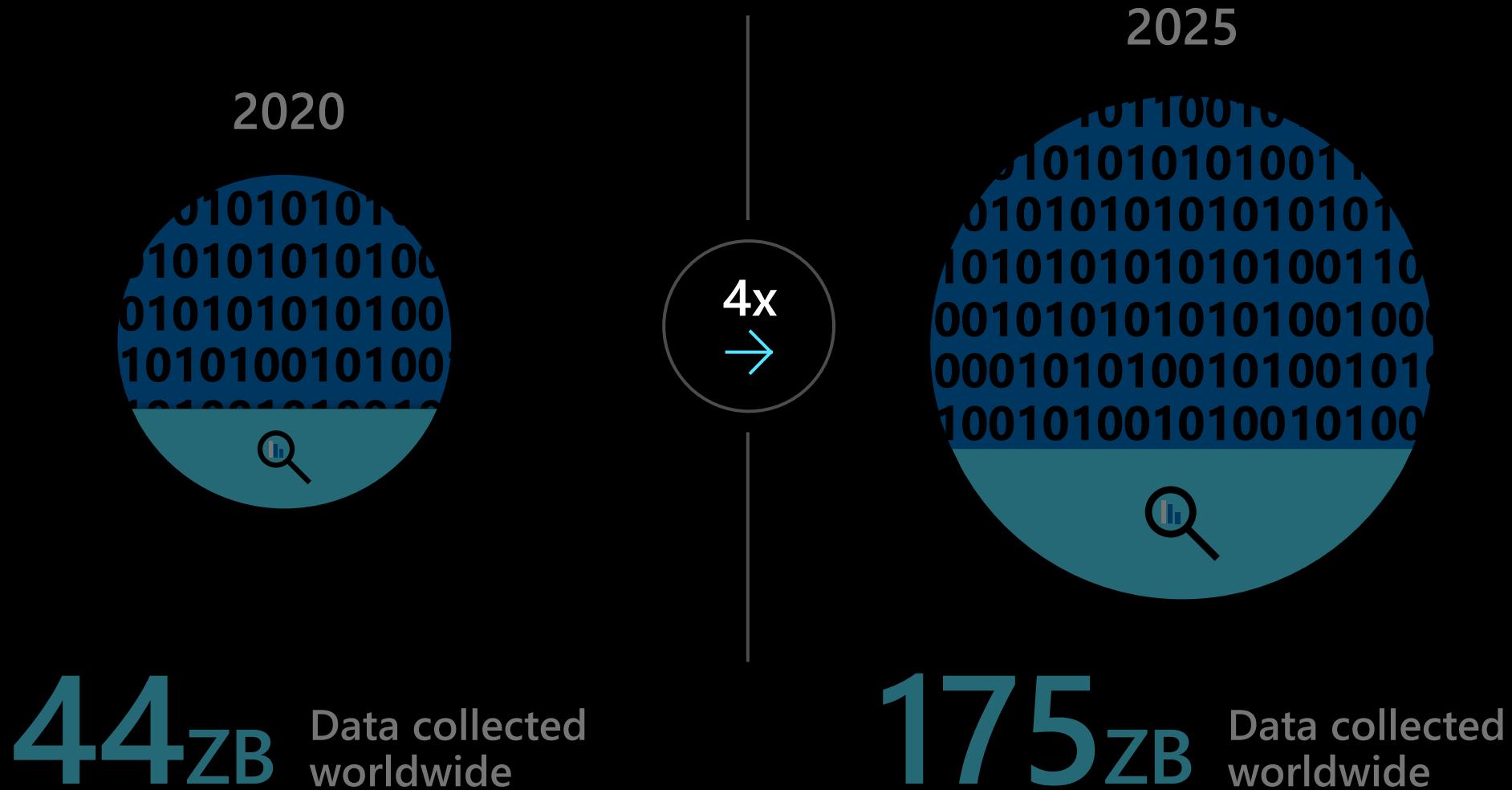
**No need to juggle
processing schedules**

**Visualization and data
science tools empower
enterprise-wide insights**

Section 3

The paradox of analytics

While data grows 400%, less than 30% gets analyzed



More analytics solutions lead to more silos



Data

- Structured
- Unstructured
- Streaming
- Big data
- Cloud
- On-premises
- IoT/edge



Technologies

- Map Reduce
- Open source projects
- Data mart
- Data warehouse
- Data lake
- Spreadsheets
- RDMS
- Data visualization tools
- ML models
- AI services



Skills

- SQL
- Python
- Java
- R
- Industry Schemas
- Data modeling
- Data cleaning
- Data integrations

This leads to the paradox of analytics

Each new technology creates another siloed operation

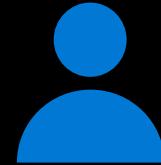
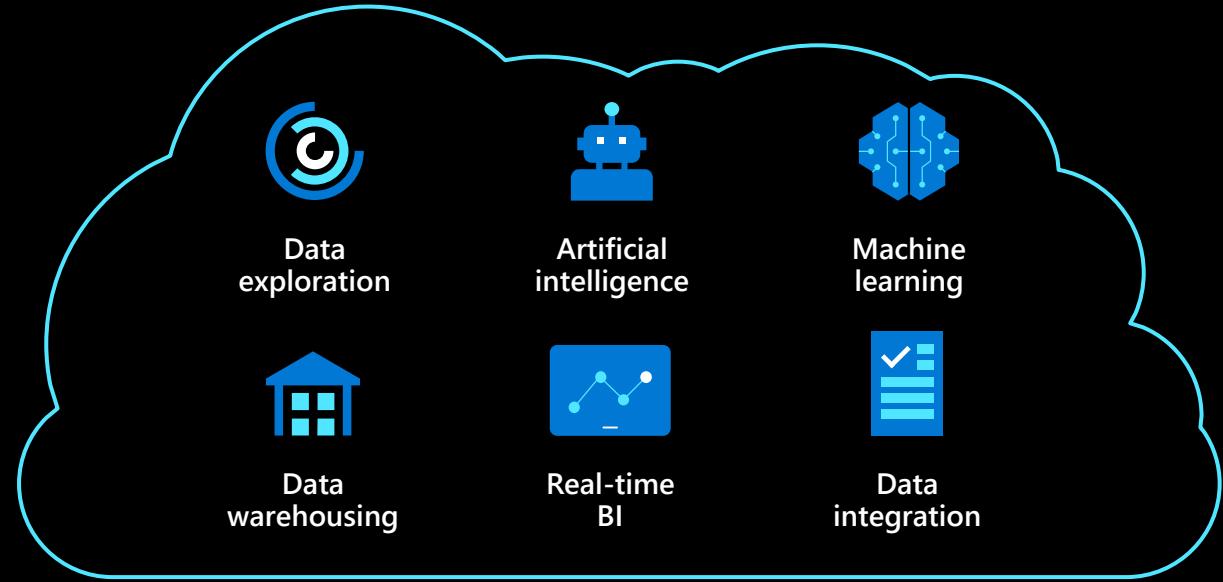
- Big data
- Data integration
- Machine learning
- Business intelligence
- Data governance
- Security



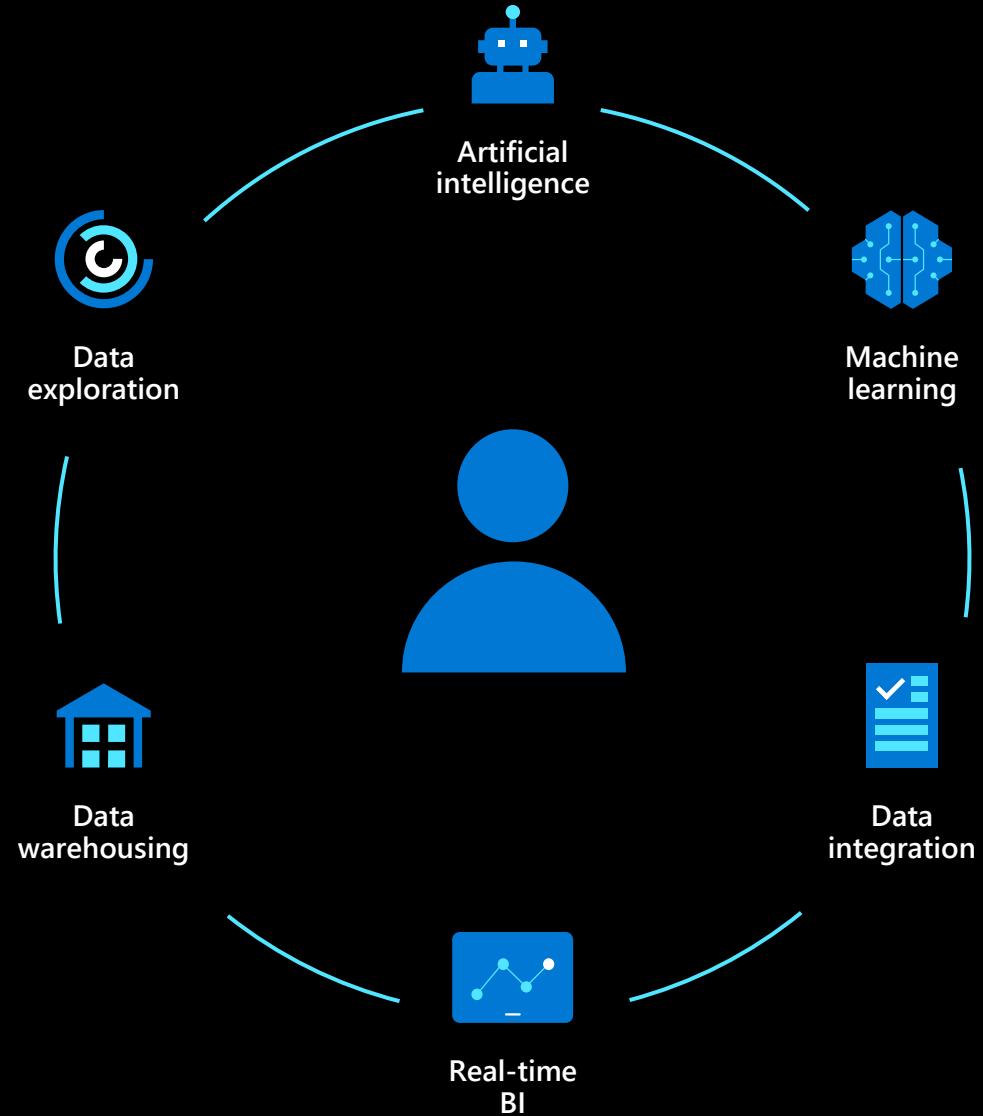
Section 4

The analytics continuum

Data professionals
shouldn't have to
continually learn new
skills to deliver insights



**Analytics should
seamlessly be part of
the way users work—
rather than a factor for
creating more siloes**

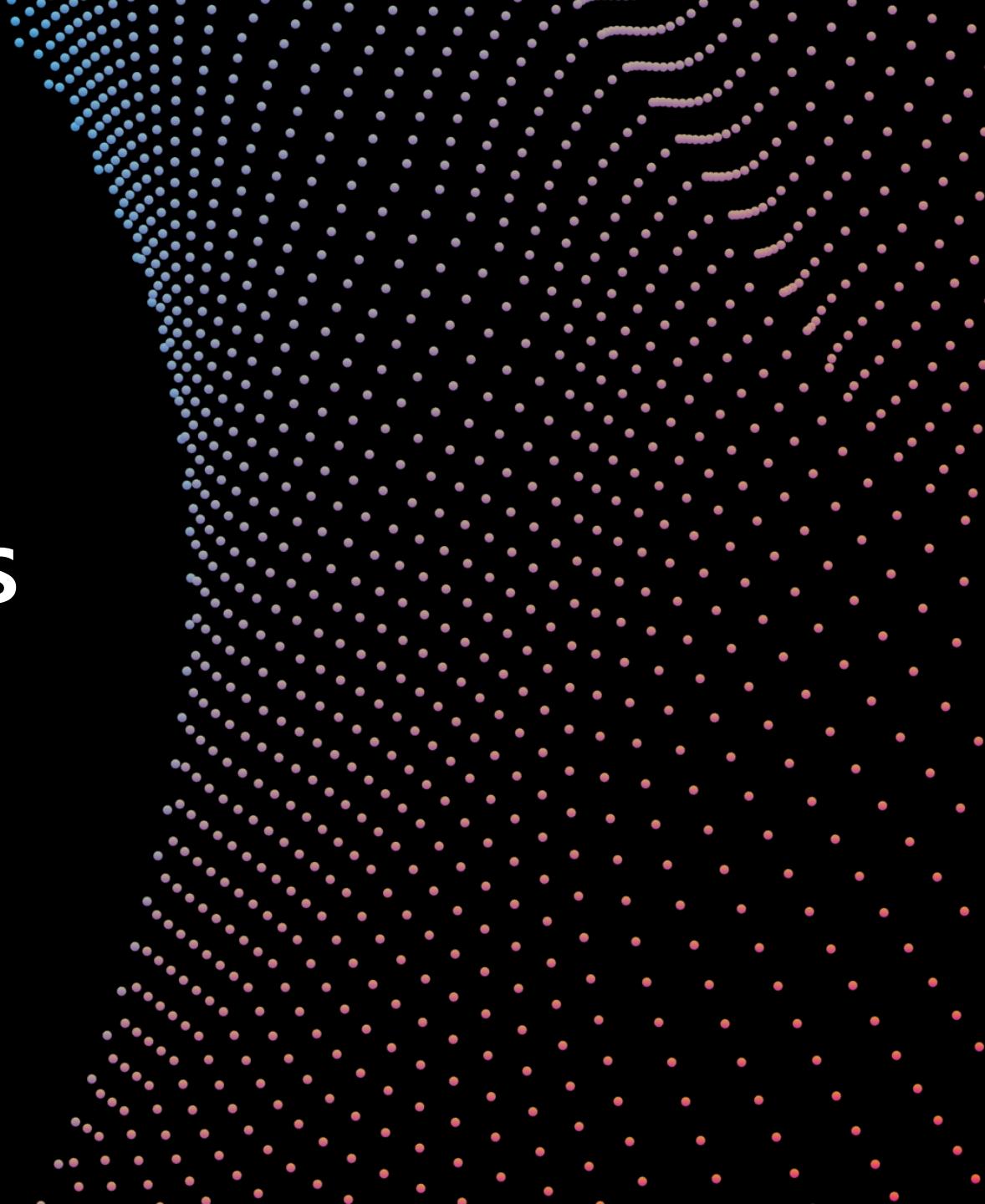


Azure Synapse
delivers the full
continuum of analytics
in a single service



Azure Synapse Analytics

**Limitless analytics service with
unmatched time to insight**



Azure Synapse Analytics

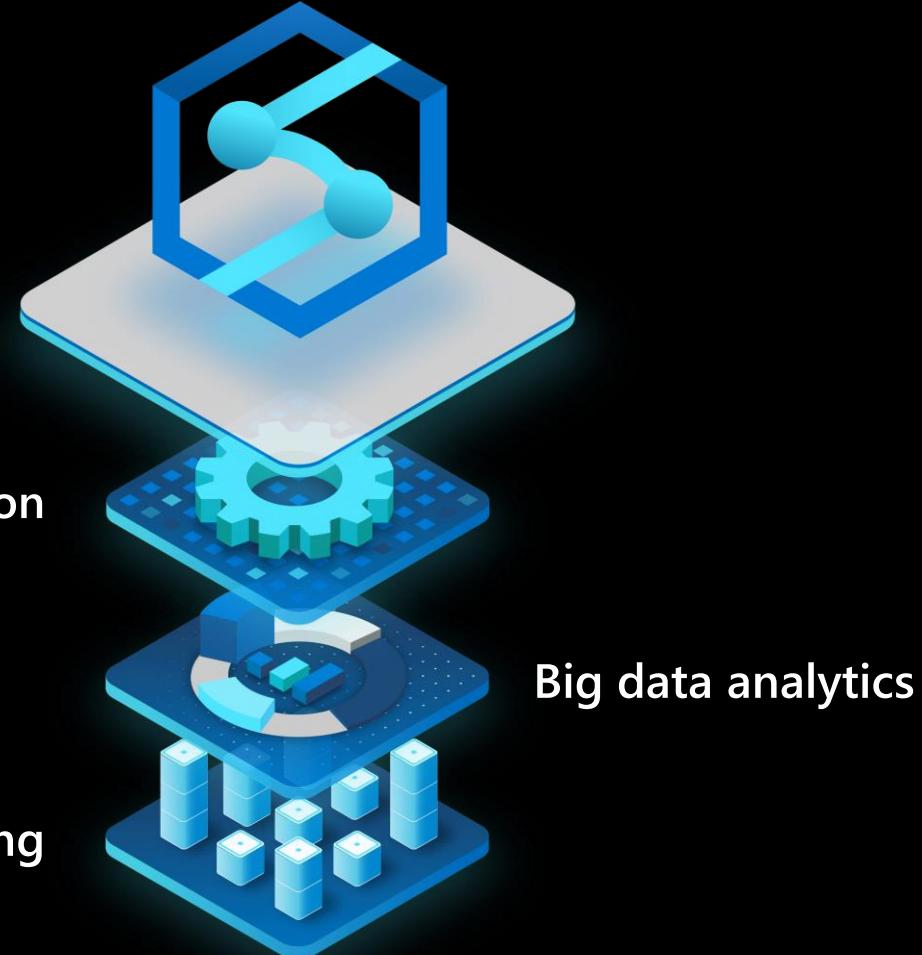
The first unified, cloud native platform for converged analytics

Azure Synapse is the only unified platform for analytics, blending big data, data warehousing, and data integration into a **single cloud native service** for end-to-end analytics at cloud scale.

Data integration

Data warehousing

Big data analytics



Azure Synapse Analytics

Powered by a new cloud native
distributed SQL engine

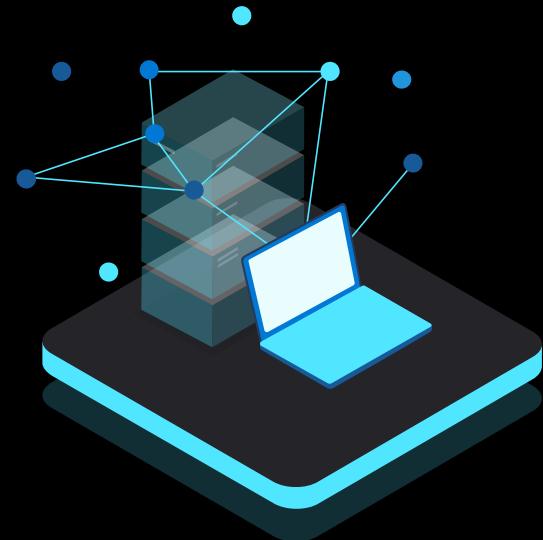


Serverless + dedicated SQL

Flexible consumption models

Serverless pay-per-query ideal for ad-hoc data lake exploration and transformation

Dedicated clusters optimized mission-critical data warehouse workloads



Serverless



Dedicated

Let's dive deeper

- Comprehensive security and compliance
- Streamlined data integration
- Flexible data warehousing
- Real-time operational analytics
- Integrated machine learning
- Power BI + Azure Synapse



Security and compliance

Unified governance controls

Complete data protection

Best-in-class security

Customer & System Managed Keys

All data encrypted by default

Up to 3x levels of data encryption at rest

Democratize data at scale with fine-grained ACL

Proactive protection

Comprehensive Compliance

Category	Feature	
Data Protection	Data in transit	✓
	Data encryption at rest	✓
	Data discovery and classification	✓
Access Control	Object level security (tables/views)	✓
	Row level security	✓
	Column level security	✓
	Dynamic data masking	✓
	Column level encryption	✓
Authentication	SQL login	✓
	Azure active directory	✓
	Multi-factor authentication	✓
Network Security	Managed virtual network	✓
	Custom virtual network	✓
	Firewall	✓
	Azure ExpressRoute	✓
	Azure Private Link	✓
Threat protection	Threat detection	✓
	Auditing	✓
	Vulnerability assessment	✓
Isolation	Dedicated metadata store	✓
	Hosted in customer tenant	✓

Managed virtual networks

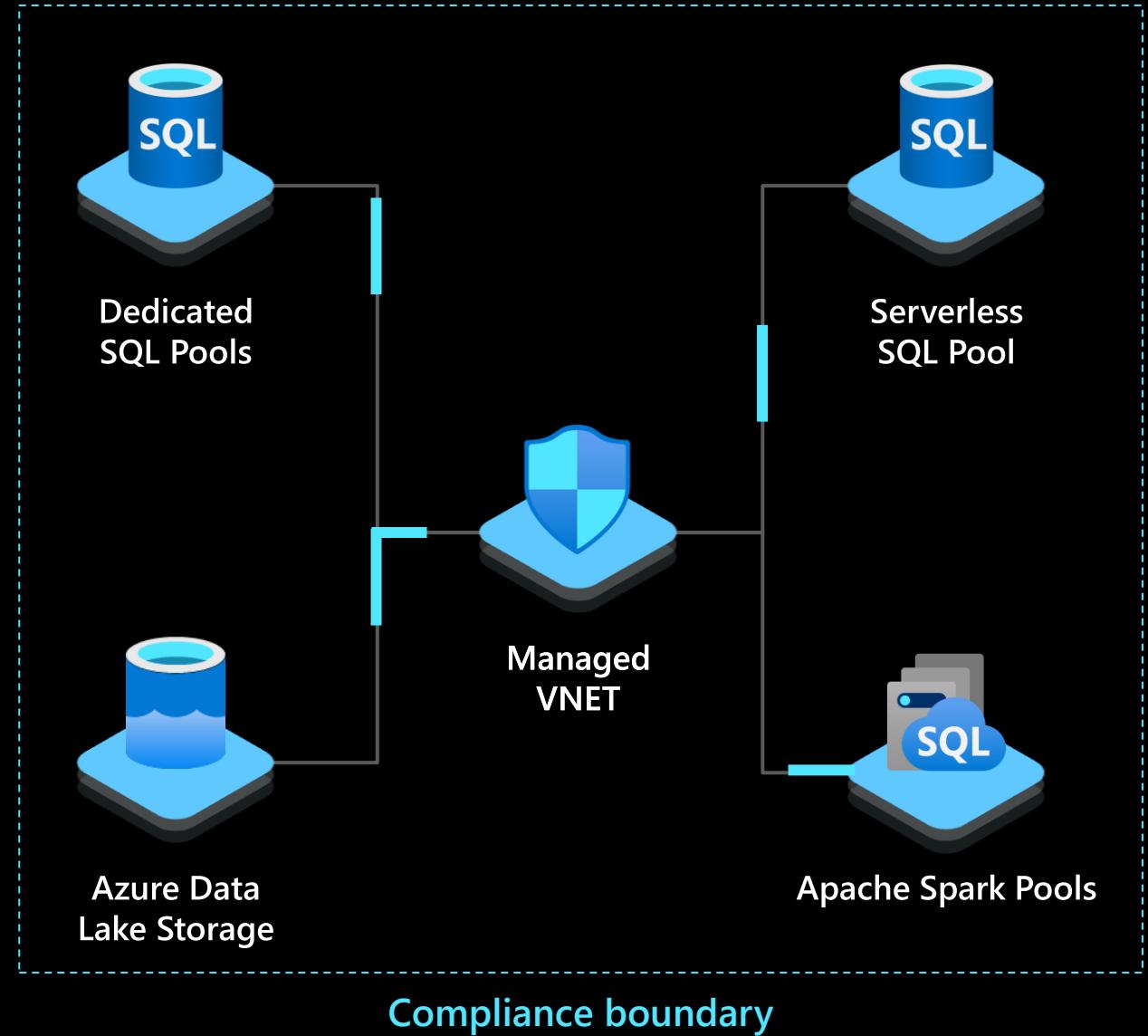
Eliminate network maintenance

One-click enables automated management of virtual networks between cluster endpoints

Synapse resources only ever interop with private endpoints

No management of subnets or IP Ranges

Prevents data exfiltration



Data integration

Code-free hybrid data integration

Hybrid data integration

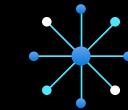
Cloud native ETL/ELT

95+ connectors available

Secure connectivity to on-premise data sources, other clouds, and SaaS applications

Code-first and low/no code design interfaces

Schedule and Event based triggering



95 connectors



All your data
(on-prem,
SaaS applications,
other clouds)

Code-free

Code-free data wrangling

No/low-code data transformation

Excel-like interface is familiar to users

Transform data to desired shape completely visually

Operationalize into pipelines

The screenshot shows the Microsoft Azure Synapse Analytics Power Query Editor interface. The top navigation bar includes 'Microsoft Azure', 'Synapse Analytics', 'wsazuresynapseanalytics', and a 'Search resources' bar. Below the navigation is a toolbar with various icons for data operations like 'Validate all', 'Publish all', and 'Discard all'. The main workspace is titled 'PQSalesPrep' under the 'Settings' tab. A message states: 'Currently not all Power Query M functions are supported for data wrangling despite being available during authoring. Learn more'. The interface features a ribbon with 'Home', 'Transform', 'Add column', and 'View' tabs. The 'Transform' tab is selected, showing a set of tools for data manipulation. On the left, there's a 'Queries' pane listing 'ADFResource' (1 item) and 'UserQuery'. The main area displays a table with 17 rows of data, each containing columns for 'ab storeId', 'ab productCode', '12 quantity', '1.2 logQuantity', 'ab advertising', 'ab price', 'ab weekStarting', and 'ab id'. The table has a total of 8 columns and 99+ rows. The bottom of the table shows the status 'Columns: 8 Rows: 99+'. The right side of the interface contains a large number of additional transformation tools and options.

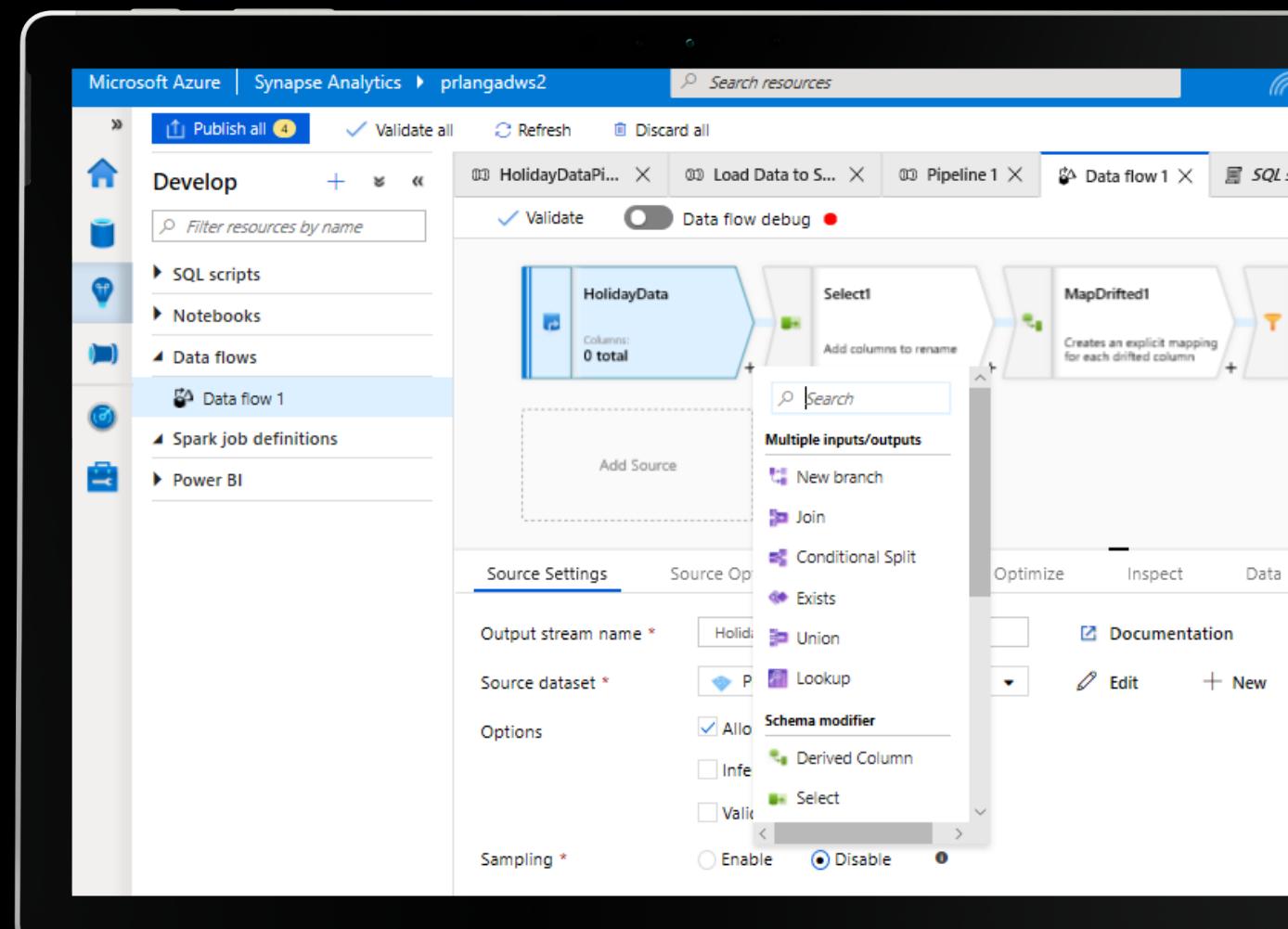
Hybrid data integration

Real-time operational analytics

No data integration pipelines required

Near-zero impact on operational systems

Latency <90s at 99th percentile



Data warehouse

Scalable and secure analytics platform for SQL workloads

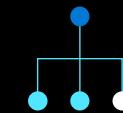
Suitable for any scale

Fully-managed elastic platform

Elastic compute that can be easily optimized to different classes of workload

All features available in a single tier

Infinite cost effective PAYG storage



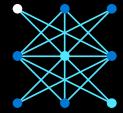
Department

Gigabytes of data
10's of users
Weekday usage



Organization

Terabytes of data
100's of users
24/7 usage



Enterprise

Petabytes of data
1000's of users
Tier-0 Availability

Synapse Studio

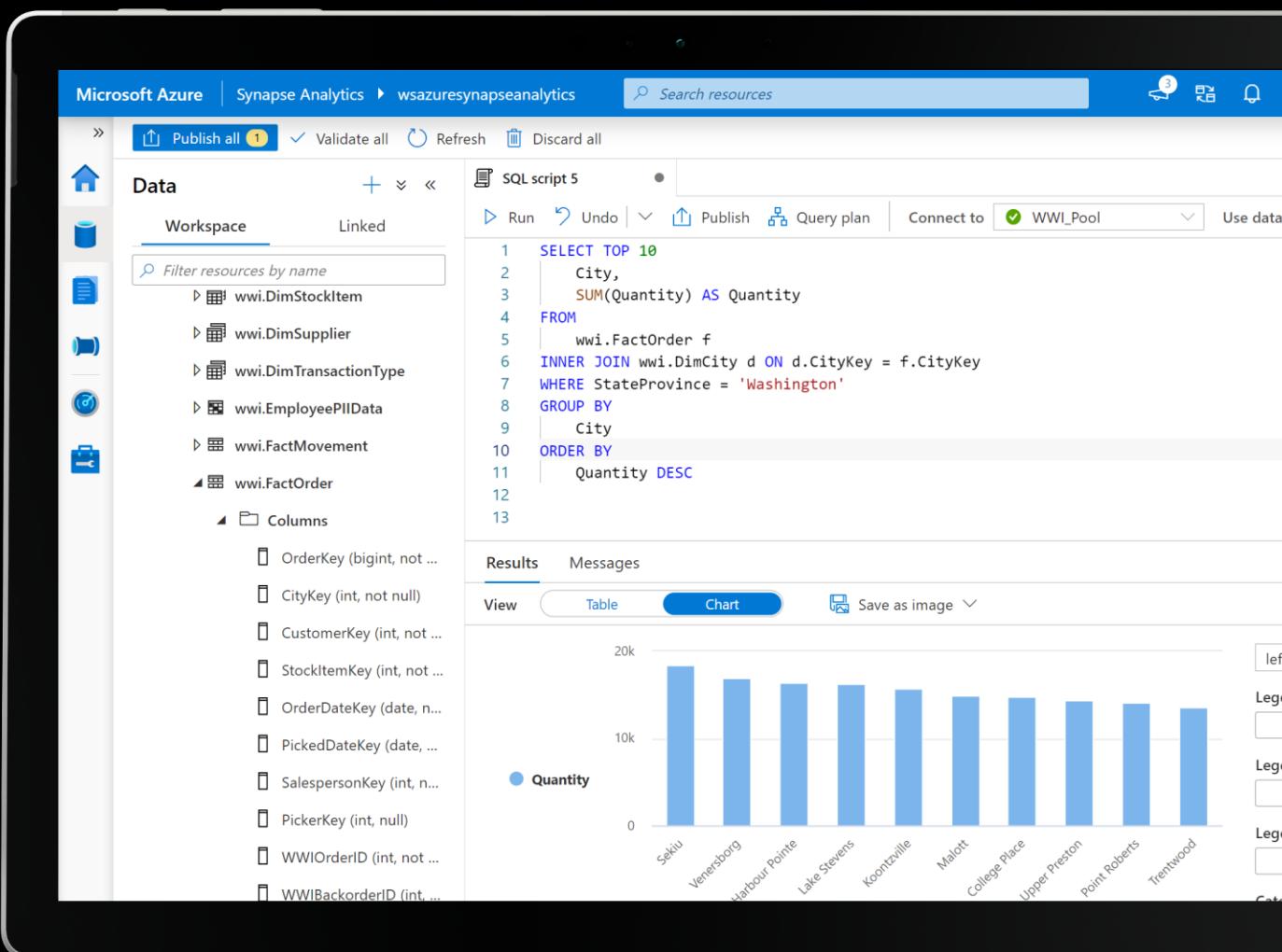
SQL Editor

Automatic code completion (Intellisense)

Script collaboration within the Workspace

Built-in visualizations

Easily switch between clusters



The screenshot shows the Microsoft Azure Synapse Analytics Studio interface. On the left, there's a sidebar with icons for Home, Data, Datasets, Pipelines, and Jobs. The main area has tabs for Data, Workspace, and Linked. In the Workspace tab, there's a search bar and a list of datasets: wwi.DimStockItem, wwi.DimSupplier, wwi.DimTransactionType, wwi.EmployeePIIData, wwi.FactMovement, wwi.FactOrder, and Columns. Below the columns is a list of columns: OrderKey, CityKey, CustomerKey, StockItemKey, OrderDateKey, PickedDateKey, SalespersonKey, PickerKey, WWIOrderID, and WWIBackorderID. On the right, there's a SQL script editor with the following query:

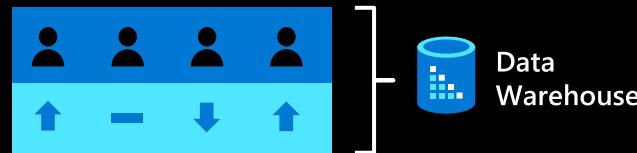
```
1 SELECT TOP 10
2     City,
3     SUM(Quantity) AS Quantity
4 FROM
5     wwi.FactOrder f
6 INNER JOIN wwi.DimCity d ON d.CityKey = f.CityKey
7 WHERE StateProvince = 'Washington'
8 GROUP BY
9     City
10 ORDER BY
11     Quantity DESC
12
13
```

Below the script are tabs for Results and Messages, with the Results tab selected. Under Results, there are View (Table, Chart), and Save as image options. A bar chart titled "Quantity" is displayed, showing the total quantity for various cities. The x-axis lists cities: Sekiu, Venerborg, Harbour Pointe, Lake Stevens, Koontzville, Malott, College Place, Upper Preston, Point Roberts, and Trentwood. The y-axis shows values from 0 to 20k.

Workload management

Scale in

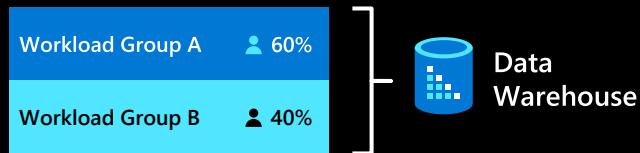
Workload Importance



Benefits:

- Predictable cost
- Bias to high-value workloads within fixed/predictable budget
- Enables customers to easily deprioritize queries which don't need to be run immediately
- Built-in starvation prevention

Workload Isolation

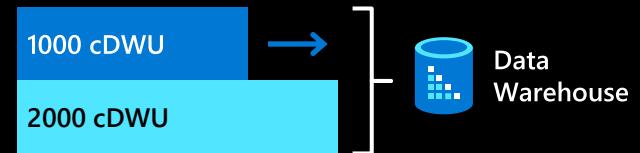


Benefits:

- Predictable cost
- Efficient for unpredictable workloads since compute can overcommit
- No cache eviction for scaling (no performance cliff at restart)
- Single connection string for isolation (unlike Snowflake virtual warehouses)

Scale out

Elastic Cluster (Scale Up)



Benefits:

- Incremental add compute
- Increase large query performance
- Single cache for heterogeneous workloads
- Single endpoint



Azure Synapse supports a more diverse set of workload management tools through workload importance, intra-cluster isolation, and elastic clusters.

Machine learning

Empower everyone with predictive insights

Machine learning

Democratize predictive power

Synapse makes predictive analytics accessible to all

Notebooks provides a code authoring experience for complex predictive models

Automatic ML graphical interface provides a no-code experience for creating ML models

Native integration with Azure Cognitive Search provides access to pre-built models



All Code

Notebook IDE
PySpark/Scala



Low/no-Code

Classification
Regression
Time-series



Pre-built models

Anomaly Detector
Sentiment Analysis



Notebook IDE code authoring

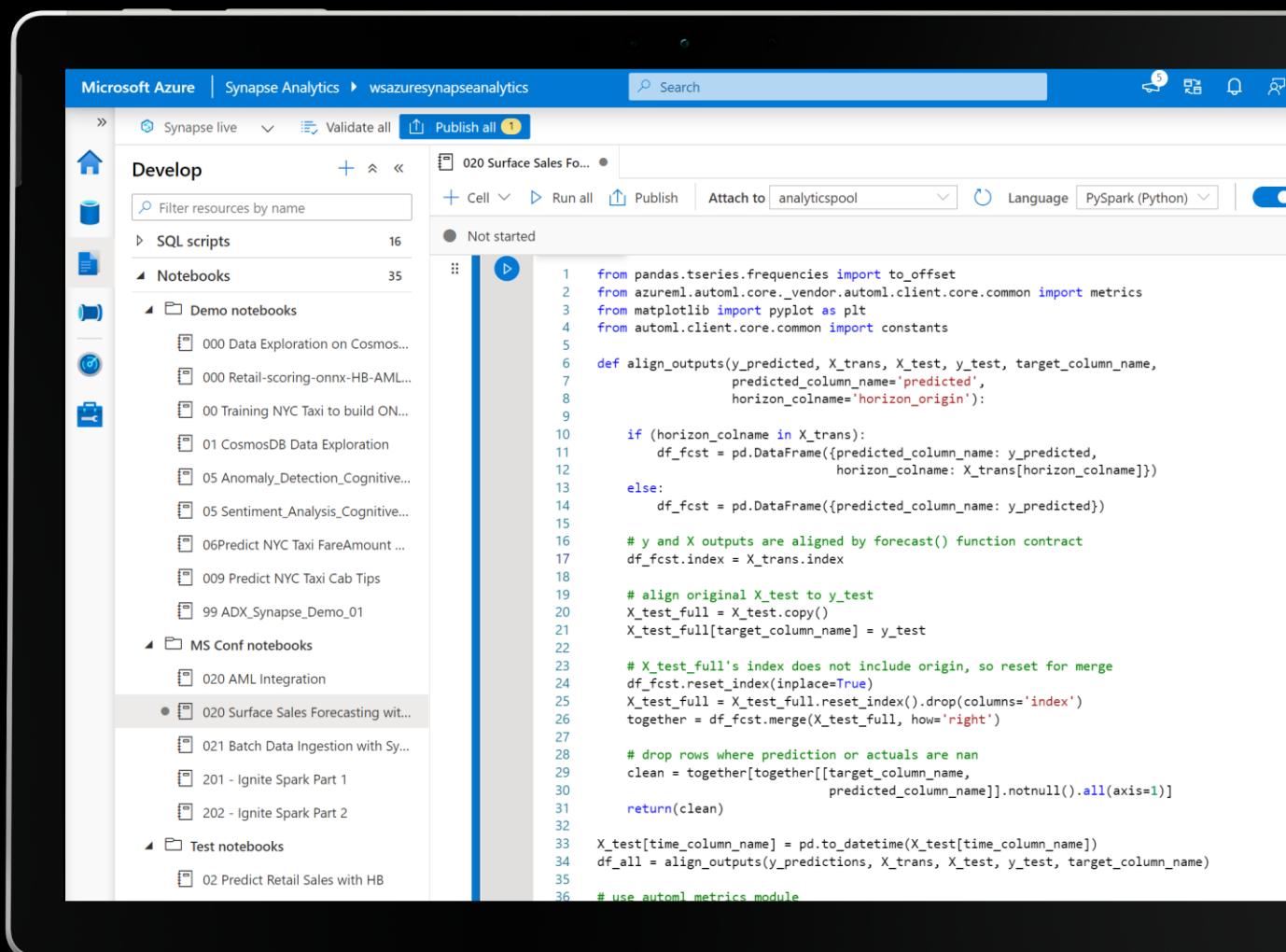
Code-first ML model development

PySpark, Scala, and C# languages supported

Automatic code completion (Intellisense)

Author multiple languages in a single notebook

Analyze data from the data warehouse, data lake, and real-time operational data from one place



The screenshot shows the Microsoft Azure Synapse Analytics Notebook IDE. The top navigation bar includes 'Microsoft Azure', 'Synapse Analytics', and a workspace name 'wsazuresynapseanalytics'. The left sidebar, titled 'Develop', lists 'SQL scripts' (16 items) and 'Notebooks' (35 items). Under 'Notebooks', there are three main categories: 'Demo notebooks' (containing 000 Data Exploration on Cosmos... through 99 ADX_Synapse_Demo_01), 'MS Conf notebooks' (containing 020 AML Integration and 020 Surface Sales Forecasting with H...), and 'Test notebooks' (containing 021 Batch Data Ingestion with Sy... and 202 - Ignite Spark Part 2). The right pane displays a code editor for a notebook titled '020 Surface Sales Fo...'. The code is written in Python and performs operations like importing pandas, aligning outputs, and merging datasets. The code editor includes line numbers, syntax highlighting, and a toolbar with 'Cell', 'Run all', 'Publish', 'Attach to', 'Language' (set to 'PySpark (Python)'), and other options.

```
from pandas.tseries.frequencies import to_offset
from azureml.core._vendor.azureml.client.core.common import metrics
from matplotlib import pyplot as plt
from automl.core.common import constants

def align_outputs(y_predicted, X_trans, X_test, y_test, target_column_name,
                  predicted_column_name='predicted',
                  horizon_colname='horizon_origin'):

    if (horizon_colname in X_trans):
        df_fcst = pd.DataFrame({predicted_column_name: y_predicted,
                               horizon_colname: X_trans[horizon_colname]})

    else:
        df_fcst = pd.DataFrame({predicted_column_name: y_predicted})

    # y and X outputs are aligned by forecast() function contract
    df_fcst.index = X_trans.index

    # align original X_test to y_test
    X_test_full = X_test.copy()
    X_test_full[target_column_name] = y_test

    # X_test_full's index does not include origin, so reset for merge
    df_fcst.reset_index(inplace=True)
    X_test_full = X_test_full.reset_index().drop(columns='index')
    together = df_fcst.merge(X_test_full, how='right')

    # drop rows where prediction or actuals are nan
    clean = together[together[[target_column_name,
                                predicted_column_name]].notnull().all(axis=1)]
    return(clean)

X_test[time_column_name] = pd.to_datetime(X_test[time_column_name])
df_all = align_outputs(y_predictions, X_trans, X_test, y_test, target_column_name)

# use automl.metrics module
```

Open ecosystem with support for industry standards

Data + Languages

Languages such as SQL, PySpark, Scala and C# in support of data science and data warehouse workloads

The data lake supports an unlimited set of file formats including Parquet, ORC and JSON as well as audio, image, and video formats

Language



C#

SQL

Data



.ORC

< , >

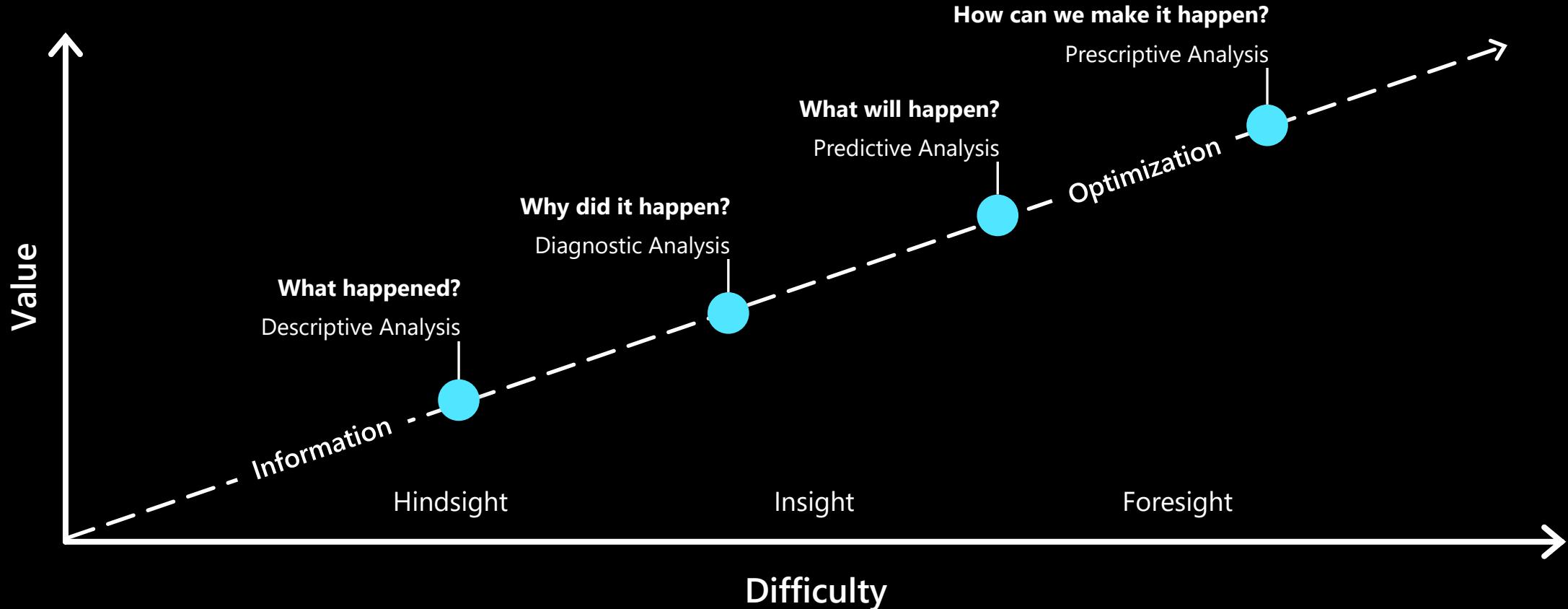
{JSON}



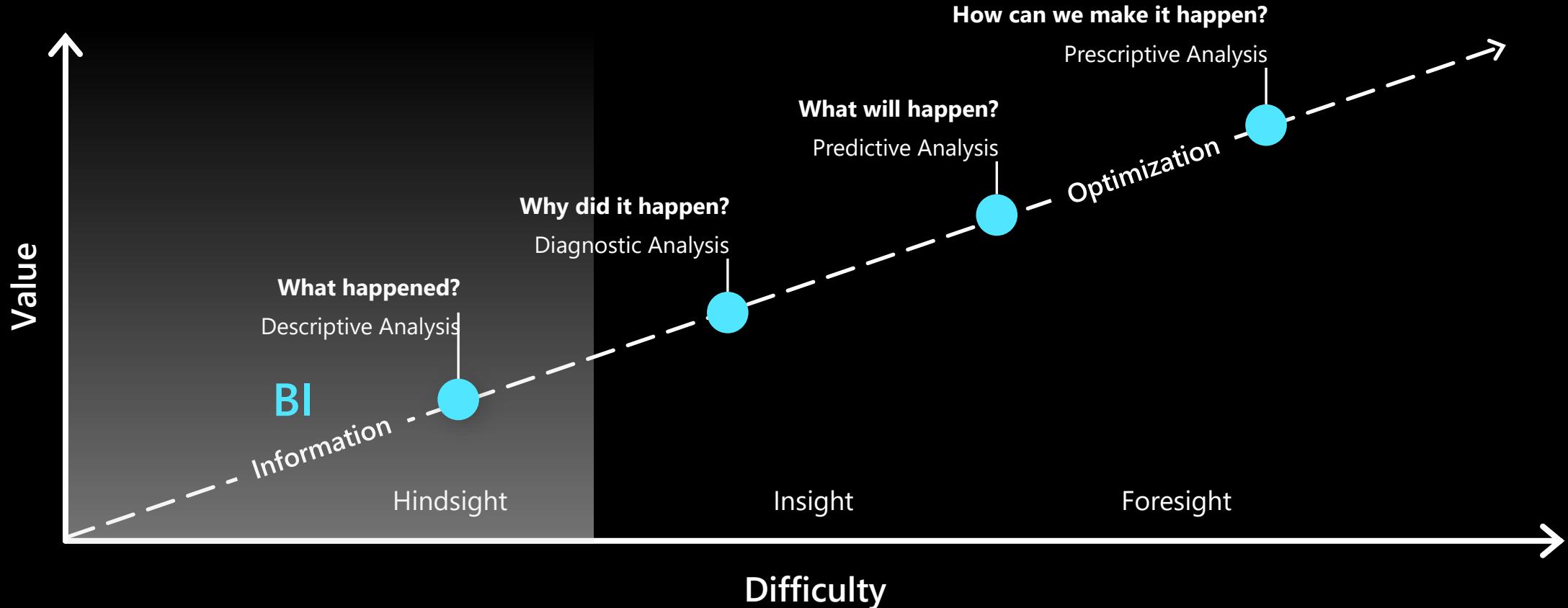
Power BI + Azure Synapse

An unmatched combination

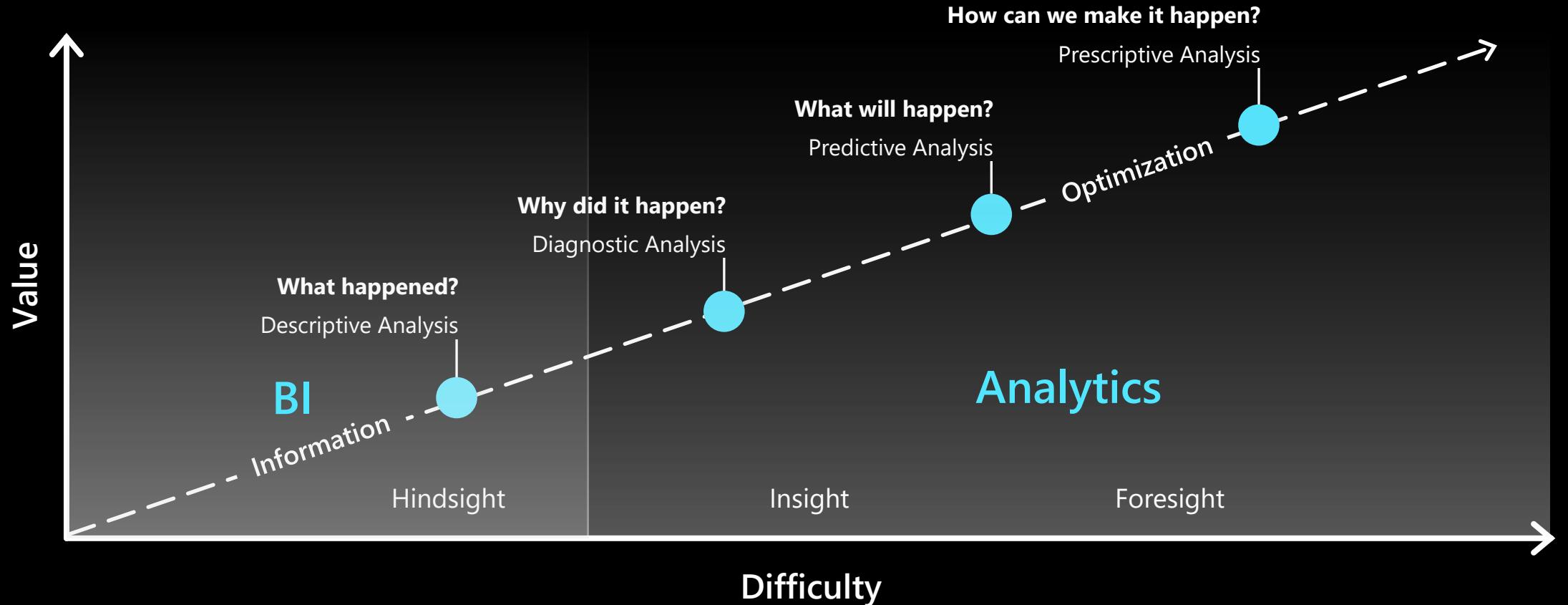
Where do you find yourself on the curve?

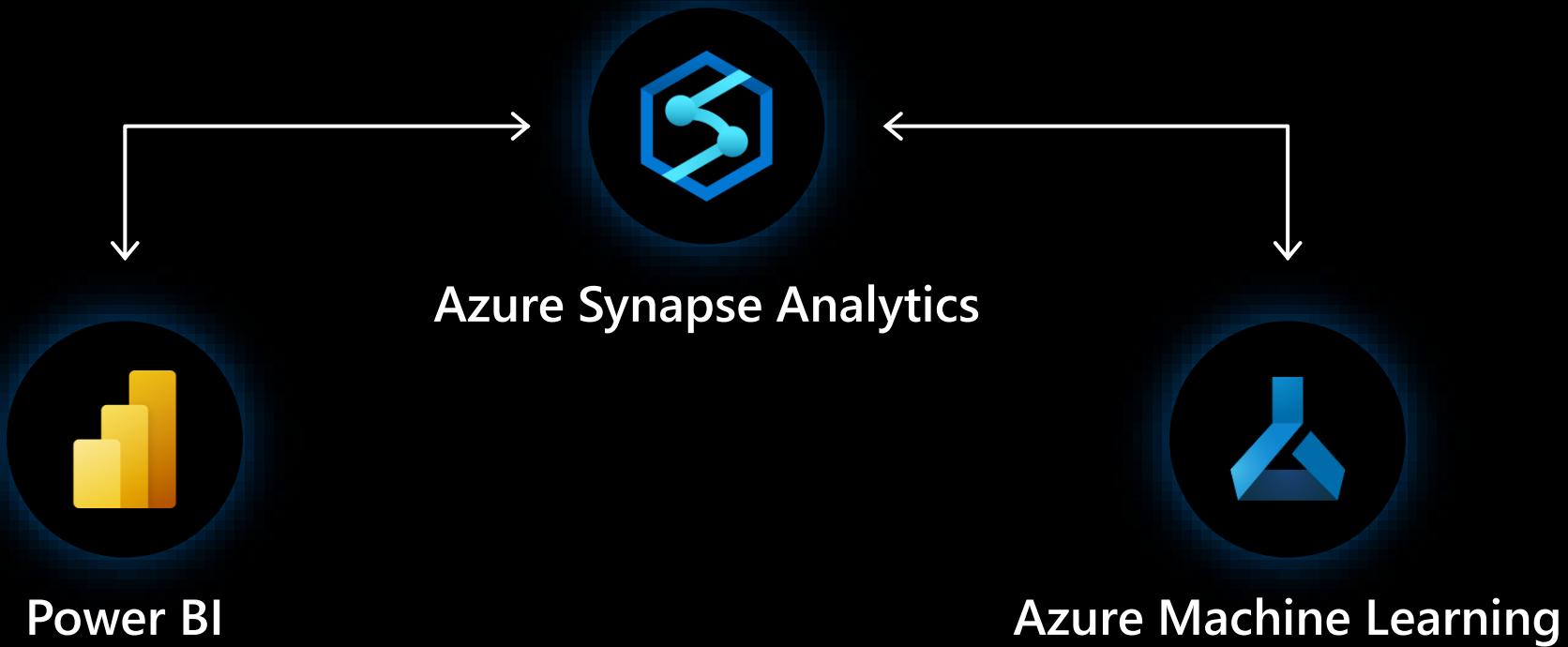


Where do you find yourself on the curve?



BI + Analytics unlock the door to AI, machine learning, and real-time insights





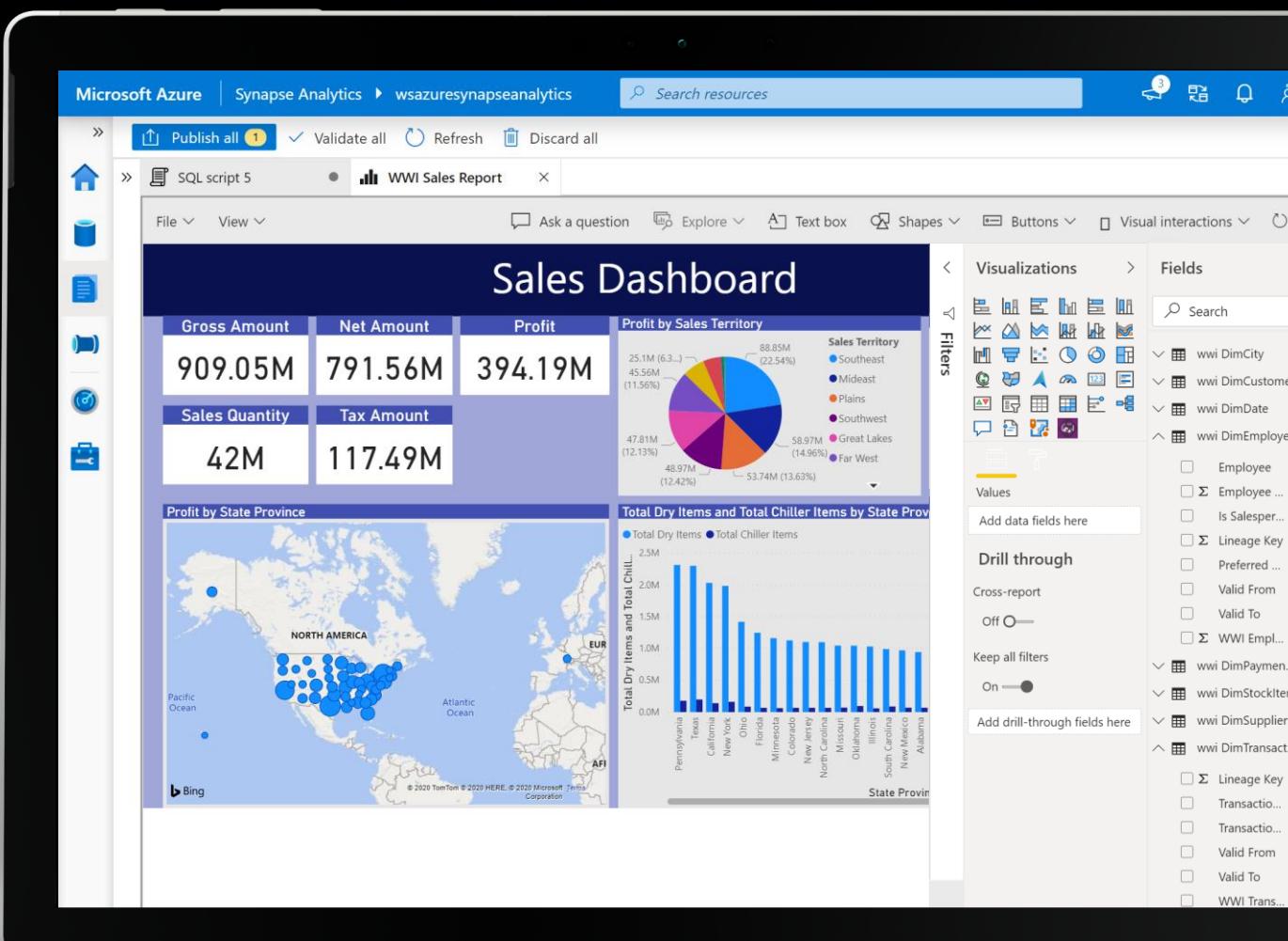
Unified experience to enrich data
Automated ML for rapid development
Seamless collaboration

Power BI integration

Build dashboard in Synapse Studio

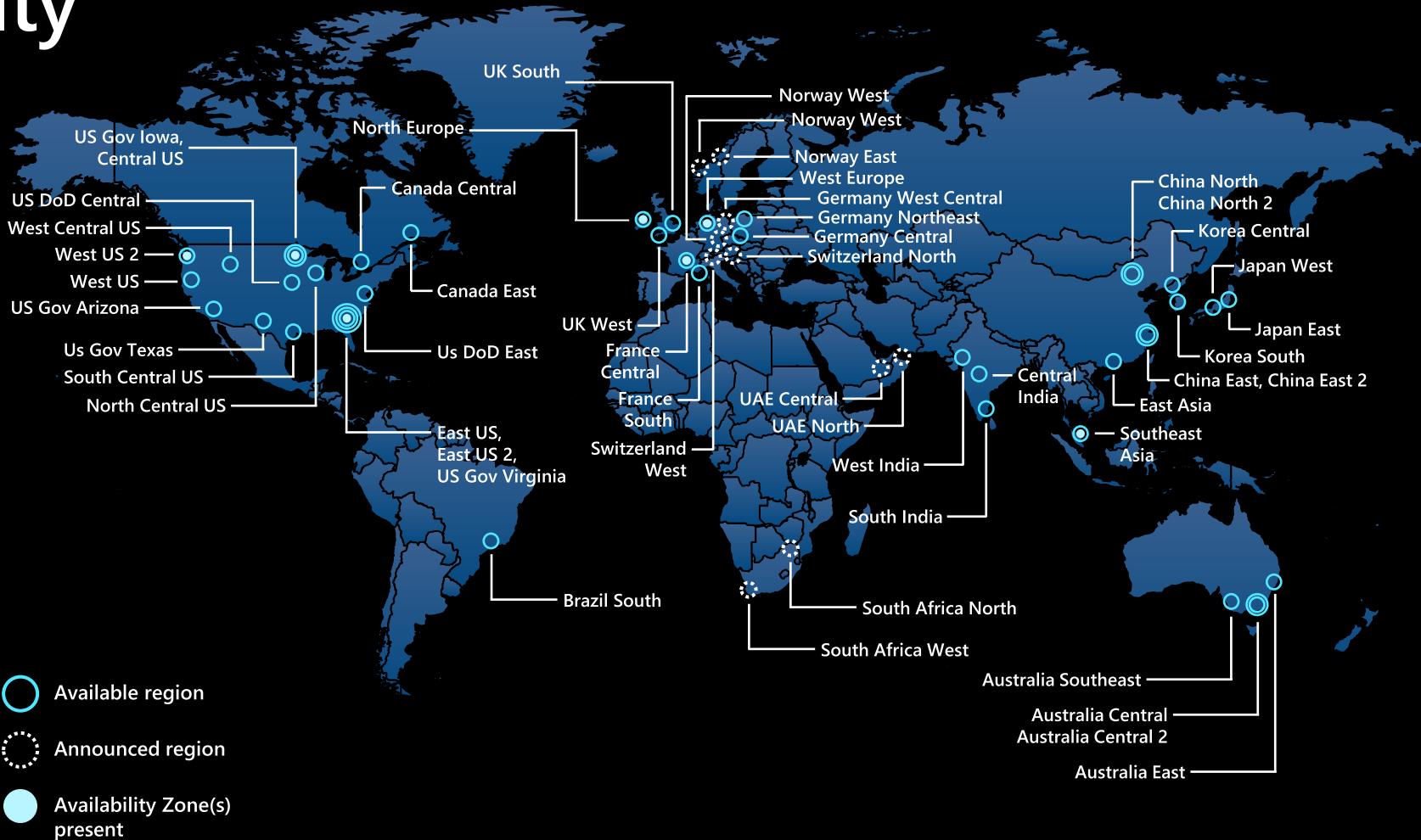
Code-free experience for development rich visualizations

One-click publishing to for secure consumption across the enterprise



Azure Synapse regional availability

Australia Southeast	Korea Central
Australia East	North Central US
Brazil South	North Europe
Canada Central	South Africa North
Canada East	South Central US
Central India	Southeast Asia
Central US	Switzerland North
East Asia	UK West
East US	UK South
East US 2	West Central US
France Central	West Europe
Germany West Central	West US
Japan East	West US 2
Japan West	



Knowledge Center

Accelerate time to solution

Azure Open Data sets

Pre-built samples to accelerate development

- SQL Scripts
- Notebooks
- Data Pipelines

The screenshot shows the Microsoft Azure Synapse Analytics Sample center interface. The top navigation bar includes 'Microsoft Azure' and 'Synapse Analytics' with a path to 'wsazuresynapseanalytics'. Below the navigation is a sidebar with icons for Home, Datasets, Notebooks, SQL scripts, and Pipelines. The main area is titled 'Sample center' and has tabs for 'Datasets', 'Notebooks', 'SQL scripts', and 'Pipelines'. A search bar says 'Filter by keyword' and a tag filter says 'Tags : All'. There are several dataset cards displayed in a grid:

Dataset	Description	ID	Action
Bing COVID-19 Data	Bing COVID-19 data includes confirmed, fatal, and recovered cases from all regions, updated daily.	ID: bing-covid-19-data	Sample
Boston Safety Data	Read data about 311 calls reported to the city of Boston. This dataset is stored in Parquet format and is updated daily.	ID: city_safety_boston	Sample
COVID Tracking Project	The COVID Tracking Project dataset provides the latest numbers on tests, confirmed cases, hospitalizations, and deaths.	ID: covid-tracking	Sample
Chicago Safety Data	Read data about 311 calls reported to the city of Chicago. This dataset is stored in Parquet format and is updated daily.	ID: city_safety_chicago	Sample
European Centre for Disease Prevention and Control (ECDC) Covid-19 Cases	The latest available public data on COVID-19 cases and variants.	ID: ecdc-covid-19-cases	Sample
NOAA Integrated Surface Data (ISD)	NOAA Integrated Surface Data (ISD) provides Worldwide hourly weather observations.	ID: isd	Sample
NYC Taxi & Limousine Commission - For-Hire Vehicle (FHV) trip records	The For-Hire Vehicle trip records include pickup and drop-off locations, vehicle type, and fare information.	ID: nyc_tlc_fhv	Sample
NYC Taxi & Limousine Commission - green taxi trip records	The green taxi trip records include pickup and drop-off locations, vehicle type, and fare information.	ID: nyc_tlc_green	Sample



Break



Hands-on lab

**Build an end-to-end analytics solution
with Azure Synapse & Power BI**



Azure Immersion Workshop: Analytics | 2021-03-16 | Denmark

Dear Steen Kristensen

Your **Azure Immersion Workshop: Analytics | 2021-03-16 | Denmark** On demand lab is ready.
You have 4Hrs, 0Mins to try out the lab before it expires.

Here is the [Azure](#) login info and lab guide details:

On Demand Lab: Azure Immersion Workshop: Analytics | 2021-03-16 | Denmark

Username: [@msazurelabs.onmicrosoft.com](#)

Password: [\[REDACTED\]](#)

Lab Guide: <https://github.com/solliancenet/azure-synapse-analytics-day>

Please use the below details for future use in your labs:

lab-ti-330680 :

Name	Value
uniqueId	330680
storage Account Name	asadatalake330680
labvm Admin Username	demouser
labvm Admin Password	Password.1!!
labvm DNS Name	labvm-nt3ddkcngslwk.westeurope.cloudapp.azure.com

[Access Lab Now](#)

If you have any questions, please contact us at cloudlabs-support@spektrasytems.com

This email is sent by Spektra Systems LLC, on behalf of Microsoft.

You are receiving this message as you have registered for On Demand Lab at
<https://experience.cloudlabs.ai>

Steen Kjøng Kristensen

Senior Consultant, Business Intelligence
Accobat A/S

skp@accobat.com

<https://www.linkedin.com/in/supersteen/>



